Optimal Newton-type methods for nonconvex smooth optimization problems

Optimal Newton-type methods for nonconvex smooth optimization problems

Coralia Cartis[*], Nicholas I. M. Gould[†] and Philippe L. Toint[‡]

June 19, 2011

Abstract

We consider a general class of second-order iterations for unconstrained optimization that includes regularization and trust-region variants of Newton's method. For each method in this class, we exhibit a smooth, bounded-below objective function, whose gradient is globally Lipschitz continuous within an open convex set containing any iterates encountered and whose Hessian is $\alpha-$Hölder continuous (for given $\alpha \in [0,1]$) on the path of the iterates, for which the method in question takes at least $\lfloor \epsilon^{-(2+\alpha)/(1+\alpha)} \rfloor$ function-evaluations to generate a first iterate whose gradient is smaller than ϵ in norm. This provides a lower bound on the evaluation complexity of second-order methods in our class when applied to smooth problems satisfying our assumptions. Furthermore, for $\alpha = 1$, this lower bound is of the same order in ϵ as the upper bound on the evaluation complexity of cubic regularization, thus implying cubic regularization has optimal worst-case evaluation complexity within our class of second-order methods.

1 Introduction

Newton's method has long represented a benchmark for rapid asymptotic convergence when minimizing smooth, unconstrained objective functions [10]. It has also been efficiently safeguarded to ensure its global convergence to first- and even second-order critical points, in the presence of local nonconvexity of the objective using linesearch [18], trust-region [9] or other regularization techniques [13, 17, 1]. Many variants of these globalization techniques have been proposed. These generally retain fast local convergence under non-degeneracy assumptions, are often suitable when solving large-scale problems and sometimes allow approximate rather than true Hessians to be employed. We attempt to capture the common features of these methods in the description of the class of methods M.α below.

In this paper, we are concerned with measuring possible inefficiency of M.α methods in terms of the number of function-evaluations required to generate approximate first-order critical points

[*]School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland, UK. Email: coralia.cartis@ed.ac.uk. All three authors are grateful to the Royal Society for its support through the International Joint Project 14265.

[†]Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK. Email: nick.gould@stfc.ac.uk. This work was supported by the EPSRC grant EP/E053351/1.

[‡]Department of Mathematics, FUNDP - University of Namur, 61, rue de Bruxelles, B-5000, Namur, Belgium. Email: philippe.toint@fundp.ac.be.

of "sufficiently smooth nonconvex objectives", as we shall define in A.α below. This amounts to establishing *lower bounds* on the evaluation complexity of the methods class M.α when applied to minimizing functions in A.α.

There is a growing literature on the global evaluation complexity of first- and second-order methods for nonconvex smooth optimization problems. In particular, it is known [19], [15, p.29] that steepest-descent method with either exact or inexact linesearches takes at most $\mathcal{O}(\epsilon^{-2})$ iterations/function-evaluations to generate a gradient whose norm is at most ϵ when started from an arbitrary initial point and applied to nonconvex smooth objectives with gradients that are globally Lipschitz continuous within some open convex set containing the iterates generated. Furthermore, this bound is sharp (for inexact linesearches) [3]. Similarly, trust-region methods that ensure at least a Cauchy (steepest-descent-like) decrease on each iteration satisfy an evaluation complexity bound of the same order under identical conditions [12]. It follows that Newton's method globalized by trust-region regularization satisfies the same $\mathcal{O}(\epsilon^{-2})$ evaluation upper bound; such a bound can also be shown to be tight [3] provided additionally that the Hessian on the path of the iterates for which pure Newton steps are taken is Lipschitz continuous.

From a worst-case complexity point of view, one can do better when a cubic regularization/perturbation of the Newton direction is used [13, 17, 1]—such a method iteratively calculates step corrections by (exactly or approximately) minimizing a cubic model formed of a quadratic approximation of the objective and the cube of a weighted norm of the step. For such a method, the worst-case global complexity improves to be of order $\epsilon^{-3/2}$ [17, 2], for problems whose gradients and Hessians are Lipschitz continuous as above; this bound is also tight [3]. If instead powers between two and three are used in the regularization, then an "intermediate" worst-case complexity of $\mathcal{O}(\epsilon^{-(2+\alpha)/(1+\alpha)})$ is obtained for such variants when applied to functions with globally $\alpha-$Hölder continuous Hessian on the path of iterates, where $\alpha \in (0, 1]$ [2].

These (tight) upper bounds on the evaluation complexity of such second-order methods naturally raise the question as to whether other second-order methods might have better worst-case complexity than cubic (or similar) regularization over certain classes of sufficiently smooth functions. To attempt to answer this question, we define a general, parametrized class of methods that includes Newton's method, and that attempts to capture the essential features of globalized Newton variants we have mentioned. Our class includes for example, the algorithms discussed above as well as multiplier-adjusting types such as the Goldfeld-Quandt-Trotter approach [11]. The methods of interest take a potentially-perturbed Newton step at each iteration so long as the perturbation is "not too large" and "sufficient decrease" is obtained. The size of the perturbation allowed is simultaneously related to the parameter α defining the class of methods and the rate of the asymptotic convergence of the method. For each method in each α-parametrized class, we construct a function with globally $\alpha-$Hölder-continuous Hessian on the path of the iterates and Lipschitz continuous gradient for which the method takes precisely $\lceil \epsilon^{-(2+\alpha)/(1+\alpha)} \rceil$ function-evaluations to drive the gradient norm below ϵ. As such counts are the same order as the upper complexity bound of regularization methods, it follows that the latter methods are optimal within their respective α-class of methods. As α approaches zero, the complexity of these methods approaches that of steepest descent, while for $\alpha = 1$, we recover that of cubic regularization. We also discuss extending our examples of inefficiency to functions with bounded level sets.

The structure of the paper is as follows. Section 2 describes the parameter-dependent class

of methods and objectives of interest; Section 2.1 gives properties of the methods such as their connection to fast asymptotic rates of convergence while Section 2.2, some examples of methods covered by our general definition of the class. Section 3 introduces the examples of inefficiency of these methods, including the case of finite minimizers. Section 4 draws our conclusions.

2 A general parametrized class of methods and objectives

Our aim is to minimize a given C^2 objective function $f(x)$, $x \in \mathbb{R}^n$. We consider methods that generate sequences of iterates $\{x_k\}$ for which $\{f(x_k)\}$ is monotonically decreasing, we let

$$f_k \stackrel{\rm def}{=} f(x_k), \quad g_k \stackrel{\rm def}{=} g(x_k) \ \text{ and } \ H_k \stackrel{\rm def}{=} H(x_k).$$

where $g(x) = \nabla_x f(x)$ and $H(x) = \nabla_{xx} f(x)$, and we denote the left-most eigenvalue of any given symmetric matrix H by $\lambda_{\min}(H)$.

Let $\alpha \in [0,1]$ be a fixed parameter. We require that our methods belong to the following class of α-dependent iterations:

M.α Given some $x_0 \in \mathbb{R}^n$, let

$$x_{k+1} = x_k + s_k, \quad k \geq 0, \tag{2.1}$$

where s_k is defined by

$$(H_k + \lambda_k I)s_k = -g_k, \tag{2.2}$$

for some λ_k such that

$$\lambda_k \geq 0 \ \text{ and } \ H_k + \lambda_k I \succeq 0. \tag{2.3}$$

Furthermore, we require that no infinite steps are taken, namely

$$\|s_k\| \leq \kappa_{\rm s}, \ \text{ for some } \kappa_{\rm s} > 0 \text{ independent of } k, \tag{2.4}$$

and that the algorithm-generated 'multiplier' λ_k satisfies

$$\lambda_k + \lambda_{\min}(H_k) \leq \kappa_\lambda \max\left\{|\lambda_{\min}(H_k)|, \|g_k\|^{\frac{\alpha}{1+\alpha}}\right\}, \tag{2.5}$$

for some $\kappa_\lambda > 1$ independent of k. □

Typically, the expression (2.2) for s_k is derived by minimizing the second-order model

$$m_k(s) = f_k + g_k^T s + \tfrac{1}{2} s^T (H_k + \beta_k I)s, \ \text{ with } \beta_k \stackrel{\rm def}{=} \beta_k(s) \geq 0 \text{ and } \beta_k \leq \lambda_k \tag{2.6}$$

of $f(x_k + s)$—possibly with an explicit regularizing constraint—with the aim of obtaining a sufficient decrease of f at the new iterate $x_{k+1} = x_k + s_k$ compared to $f(x_k)$. In the definition of an M.α method however, the issue of (sufficient) objective-function decrease is not explicitly addressed/required. There is no loss of generality in doing so here since although local refinement of the model may be required to ensure function decrease, the number of function evaluations to do so (at least for known methods) does not increase the overall complexity by more than a constant multiple and thus does not affect quantitatively the worst-case bounds derived; see for example, [3, 2, 12] and also Section 2.2. Furthermore, the examples of inefficiency we demonstrate

in Section 3 are constructed in such a way that each iteration of the method would automatically count as "successful", that is, it provides (Cauchy-like) sufficient decrease of f.

Note that methods in M.1 are naturally included in M.α for any $\alpha \in [0,1]$ since the α exponent only occurs explicitly in the condition (2.5). The class M.0, corresponds to the case when λ_k is uniformly bounded above.

Having defined the classes of methods we shall be concerned with, we now specify the problem classes that we shall apply them to. Specifically, we are interested in minimizing functions f that satisfy

A.α $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and bounded below, with gradient g being globally Lipschitz continuous on $\mathcal{X}$ with constant L_g, namely,

$$\|g(x) - g(y)\| \leq L_g \|x - y\|, \text{ for all } x, y \in \mathcal{X}; \tag{2.7}$$

where $\mathcal{X}$ is an open convex set containing the intervals $[x_k, x_k + s_k]$, and the Hessian H being globally $\alpha-$Hölder continuous on the path of the iterates with constant $L_{H,\alpha}$, i.e.,

$$\|H(x) - H(x_k)\| \leq L_{H,\alpha} \|x - x_k\|^\alpha, \text{ for all } x \in [x_k, x_k + s_k] \text{ and } k \geq 0. \tag{2.8}$$

□

Note that the case when $\alpha = 1$ in A.α corresponds to the Hessian of f being globally Lipschitz continuous on the path of the iterates. Furthermore, the class of functions A.1 is included in A.α for any $\alpha \in (0,1)$. Also, in the case when $\alpha = 0$, (2.7) implies (2.8) holds, so that the A.0 class is that of twice continuously differentiable functions with globally Lipschitz continuous gradient on $\mathcal{X}$. Note also that the class A.α with $\alpha > 1$ contains only quadratic functions.

The next section provides some justification for the technical condition (2.5) by relating it to fast rates of asymptotic convergence. In Section 2.2, we illustrate some methods that belong to M.α.

2.1 Properties of the methods in M.α

We first simplify the assumption (2.5) in the definition of the class M.α by giving a sufficient, more concise, condition on the algorithm-generated λ_k that implies (2.5).

Lemma 2.1. Let (2.2) and (2.3) hold. Assume also that the algorithm-generated λ_k satisfies

$$\lambda_k \leq \overline{\kappa}_\lambda \|s_k\|^\alpha, \text{ for some } \overline{\kappa}_\lambda > 1 \text{ and } \alpha \in [0,1] \text{ independent of } k. \tag{2.9}$$

Then (2.5) holds with $\kappa_\lambda \stackrel{\text{def}}{=} 2\overline{\kappa}_\lambda^{\frac{1}{1+\alpha}}$.

Proof. Clearly, (2.5) holds when $\lambda_k + \lambda_{\min}(H_k) = 0$. When $\lambda_k + \lambda^{\min}(H_k) > 0$, and hence, $H_k + \lambda_k I \succ 0$, we have from (2.2) that

$$\|s_k\|^2 = \sum_{i=1}^n \frac{(\gamma_k^i)^2}{(\lambda_k + \lambda^i(H_k))^2},$$

where $H_k = U_k \Sigma_k U_k^T$, $\gamma_k = U_k^T g_k$ and $\Sigma_k := \mathrm{Diag}(\lambda^i(H_k))$, the eigenvalues of H_k. This straightforwardly implies the bound

$$\|s_k\| \leq \frac{\|g_k\|}{\lambda_k + \lambda_{\min}(H_k)}, \quad \text{whenever } H_k + \lambda_k I \succ 0. \tag{2.10}$$

This and (2.9) give the inequality

$$\lambda_k^{1+\frac{1}{\alpha}} + \lambda_k^{\frac{1}{\alpha}} \cdot \lambda_{\min}(H_k) - \overline{\kappa}_\lambda^{\frac{1}{\alpha}} \|g_k\| \leq 0. \tag{2.11}$$

Let us consider (2.11) as a function of $\lambda = \lambda_k$. Note that (2.11) is satisfied at $\lambda = \max\{0, -\lambda_{\min}(H_k)\}$, and that the left-hand side of (2.11) is strictly increasing for $\lambda > \max\{0, -\lambda_{\min}(H_k)\}$, due to (2.3). Thus any value $\lambda_k^* \geq \max\{0, -\lambda_{\min}(H_k)\}$ at which the left-hand side of (2.11) is positive will provide an upper bound on λ_k. Letting

$$\lambda_k^* \stackrel{\text{def}}{=} -\lambda_{\min}(H_k) + 2\max\left\{|\lambda_{\min}(H_k)|, \overline{\kappa}_\lambda^{\frac{1}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}}\right\}, \tag{2.12}$$

which clearly satisfies $\lambda_k^* \geq \max\{0, -\lambda_{\min}(H_k)\}$, it is straightforward to verify that (2.11) does not hold at $\lambda = \lambda_k^*$. Thus $\lambda_k \leq \lambda_k^*$, which due to (2.12) and $\overline{\kappa}_\lambda > 1$, implies (2.5). □

The requirement (2.9) crucially implies the following property regarding the length of the step generated by methods in M.α when applied to functions satisfying A.α.

Lemma 2.2. Assume that an objective function f satisfying A.α is minimized by a method in the class M.α for which (2.9) holds. Then there exists $\overline{\kappa}_{s,\alpha} > 0$ independent of k such that

$$\|s_k\| \geq \overline{\kappa}_{s,\alpha} \|g_{k+1}\|^{\frac{1}{1+\alpha}}, \quad k \geq 0. \tag{2.13}$$

Proof. The triangle inequality provides

$$\|g_{k+1}\| \leq \|g_{k+1} - (g_k + H_k s_k)\| + \|g_k + H_k s_k\|. \tag{2.14}$$

From (2.1), $g_{k+1} = g(x_k + s_k)$ and Taylor expansion provides $g_{k+1} = g_k + \int_0^1 H(x_k + \tau s_k) s_k d\tau$. This and (2.8) now imply

$$\|g_{k+1} - (g_k + H_k s_k)\| \leq \left\| \int_0^1 [H(x_k + \tau s_k) - H(x_k)] d\tau \right\| \cdot \|s_k\| \leq L_{H,\alpha} (1+\alpha)^{-1} \|s_k\|^{1+\alpha},$$

so that (2.14) and (2.2) together give

$$\|g_{k+1}\| \leq L_{H,\alpha} (1+\alpha)^{-1} \|s_k\|^{1+\alpha} + \lambda_k \|s_k\|. \tag{2.15}$$

Now (2.15) and (2.9) give (2.13) with $\overline{\kappa}_{s,\alpha} \stackrel{\text{def}}{=} [L_{H,\alpha}(1+\alpha)^{-1} + \overline{\kappa}_\lambda]^{-\frac{1}{1+\alpha}}$. □

Next, we show that (2.13) is a necessary condition for fast local convergence of methods of type (2.2), under reasonable assumptions; fast local rate of convergence in a neighbourhood of well-behaved minimizers is a "trademark" of what is commonly regarded as second-order methods.

Lemma 2.3. Let f satisfy assumptions A.α. Apply an algorithm to minimizing f that satisfies (2.1) and (2.2) and for which

$$\lambda_k \leq \overline{\kappa}_\lambda, \quad k \geq 0, \quad \text{for some } \overline{\kappa}_\lambda > 0 \text{ independent of } k. \tag{2.16}$$

Assume also that convergence at linear or faster than linear rate occurs, namely,

$$\|g_{k+1}\| \leq \kappa_c \|g_k\|^{1+\alpha}, \quad k \geq 0, \tag{2.17}$$

for some $\kappa_c > 0$ independent of k, with $\kappa_c \in (0,1)$ when $\alpha = 0$. Then (2.13) holds.

Proof. Let

$$0 \leq \alpha_k \stackrel{\text{def}}{=} \frac{\|s_k\|}{\|g_{k+1}\|^{\frac{1}{1+\alpha}}}, \quad k \geq 0. \tag{2.18}$$

From (2.17) and the definition of α_k in (2.18), we have that

$$\frac{\|s_k\|}{\alpha_k} \leq \kappa_{c,\alpha}\|g_k\| = \kappa_{c,\alpha}\|(H_k + \lambda_k I)s_k\| \leq \kappa_{c,\alpha}\|H_k + \lambda_k I\| \cdot \|s_k\|, \quad k \geq 0,$$

where $\kappa_{c,\alpha} \stackrel{\text{def}}{=} \kappa_c^{\frac{1}{1+\alpha}}$ and where we used (2.2) to obtain the first equality. It follows that

$$\|H_k + \lambda_k I\| \geq \frac{1}{\alpha_k \kappa_{c,\alpha}}, \quad k \geq 0. \tag{2.19}$$

As g is globally Lipschitz continuous in $\mathcal{X}$ due to A.α, we have that $\{H_k\}$ is bounded above for $k \geq 0$ [15, Lemma 1.2.2]. This and (2.16) imply that $\{H_k + \lambda_k I\}$ is uniformly bounded above for all k, namely,

$$\|H_k + \lambda_k I\| \leq \kappa_{hl}, \quad k \geq 0, \tag{2.20}$$

where $\kappa_{hl} \stackrel{\text{def}}{=} L_g + \overline{\kappa}_\lambda$. Now (2.19) and (2.20) give that $\alpha_k \geq 1/(\kappa_{hl}\kappa_{c,\alpha}) > 0$, for all $k \geq 0$, and so it follows from (2.18), that (2.13) holds with $\overline{\kappa}_{s,\alpha} \stackrel{\text{def}}{=} 1/(\kappa_{c_1}\kappa_{c,\alpha})$. □

It is clear from the proof of Lemma 2.3 that (2.17) is only needed asymptotically, that is for all k sufficiently large; for simplicity, we have assumed it holds globally.

Note that letting $\alpha = 1$ in Lemma 2.3 provides a necessary condition for quadratically convergent methods satisfying (2.1), (2.2) and (2.16). Also, similarly to the above proof, one can show that if superlinear convergence of $\{g_k\}$ to zero occurs, then (2.13) holds with $\alpha = 0$ for all $\overline{\kappa}_{s,\alpha} > 0$, or equivalently, $\|g_{k+1}\|/\|s_k\| \to 0$, as $k \to \infty$.

2.2 Some examples of methods that belong to the class M.α

Let us now illustrate some of the methods that either by construction or under certain conditions belong to M.α. This list of methods does not attempt to be exhaustive and other practical methods may be found to belong to M.α.

Newton's method [10]. Newton's method for convex optimization is characterised by finding a correction s_k that satisfies

$$H_k s_k = -g_k.$$

Letting

$$\lambda_k = 0 \ \text{ and } \ \beta_k = 0 \tag{2.21}$$

in (2.2) and (2.6), respectively, yields Newton's method. Provided additionally that both $g_k \in$ Range(H_k) and H_k is positive semi-definite, s_k is a descent direction and (2.3) holds. Since (2.5) is trivially satisfied in this case, it follows that Newton's method belongs to the class M.α, for any $\alpha \in [0,1]$, provided it does not generate infinite steps to violate (2.4). As Newton's method is commonly embedded within trust-region or regularization frameworks when applied to nonconvex functions, (2.4) will in fact, hold as it is generally enforced for the latter methods as shown below.

It is known [3] that Newton's method may take at least ϵ^{-2} function-evaluations to generate $\|g_k\| \leq \epsilon$ when applied to f in A.1. Here, we show that Newton's method can take at least $\epsilon^{-(2+\alpha)/(1+\alpha)}$ evaluations when applied to a function f in A.α.

Regularization algorithms [13, 15, 2]. In these methods, the step s_k from the current iterate x_k is computed by globally minimizing the model

$$m_k(s) = f_k + g_k^T s + \tfrac{1}{2} s^T H_k s + \frac{\sigma_k}{2+\alpha}\|s\|^{2+\alpha}, \tag{2.22}$$

where the regularization weight σ_k is adjusted to ensure sufficient decrease of f at $x_k + s_k$. The scalar α is the same fixed parameter as in the definition of A.α and M.α, so that for each $\alpha \in [0,1]$, we have a different regularization term and hence what we shall call an $(2+\alpha)$-regularization method. For $\alpha = 1$, we recover the cubic regularization (ARC) approach [13, 17, 1, 2]. For $\alpha = 0$, we obtain a quadratic regularization scheme, reminiscent of the Levenberg-Morrison-Marquardt method [18]. For these $(2+\alpha)$-regularization methods, we have

$$\alpha \in [0,1], \quad \lambda_k = \sigma_k \|s_k\|^\alpha \ \text{ and } \ \beta_k = \frac{2}{2+\alpha}\sigma_k \|s_k\|^\alpha \tag{2.23}$$

in (2.2) and (2.6). Since $\alpha \geq 0$, we have $0 \leq \beta_k \leq \lambda_k$ which is required in (2.6). A mechanism of successful and unsuccessful iterations and σ_k adjustments can be devised similarly to ARC [1, Alg. 2.1] in order to deal with steps s_k that do not give sufficient decrease in the objective. An upper bound on the number of unsuccessful iterations which is constant multiple of successful ones can be given under mild assumptions on f [2, Thm. 2.1]. Note that each (successful or unsuccessful) iteration requires one function- and at most one gradient-evaluation.

We now show that for each $\alpha \in [0,1]$, the $(2+\alpha)-$regularization method based on the model (2.22) satisfies (2.4) and (2.5) when applied to f in A.α, and so it belongs to M.α.

Lemma 2.4. Let f satisfy A.α with $\alpha \in (0,1]$. Consider minimizing f by applying an $(2+\alpha)$-regularization method based on the model (2.22), where the step s_k is chosen as the global minimizer of the local $\alpha-$model, namely of $m_k(s)$ in (2.6) with the choice (2.23), and where the regularization parameter σ_k is chosen to ensure that

$$\sigma_k \geq \sigma_{\min}, \quad k \geq 0, \tag{2.24}$$

for some $\sigma_{\min} > 0$ independent of k. Then (2.4) and (2.9) hold, and so the $(2+\alpha)$-regularization method belongs to M.α. Furthermore, the method requires at most

$$\left\lceil \kappa_r \epsilon^{-\frac{2+\alpha}{1+\alpha}} \right\rceil \tag{2.25}$$

function evaluations to generate $\|g_k\| \leq \epsilon$, where $\kappa_r \stackrel{\text{def}}{=} (1+\kappa_r^U)(2+\kappa_r^S)$ with $\kappa_r^U \stackrel{\text{def}}{=} c_U \log(\sigma_{\max}/\sigma_{\min})$, $\kappa_r^S \stackrel{\text{def}}{=} c_S(f(x_0)-f_{\low})/(\sigma_{\min}\sigma_{\max}^{(2+\alpha)/(1+\alpha)})$, $\sigma_{\max} \stackrel{\text{def}}{=} c_\sigma \max(\sigma_0, L_{H,\alpha})$; $f_{\low}$ is some lower bound on $\{f(x_k)\}$, while c_U, c_S and c_σ are constants depending solely on α and algorithm parameters.

Proof. The same argument that is used in [1, Lem.2.2] (for the $\alpha = 1$ case) provides

$$\|s_k\| \leq \max\left\{\left(\frac{3(2+\alpha)L_g}{4\sigma_k}\right)^{\frac{1}{\alpha}}, \left(\frac{3(2+\alpha)\|g_k\|}{\sigma_k}\right)^{\frac{1}{1+\alpha}}\right\}, \quad k \geq 0,$$

so long as A.α holds, which together with (2.24), implies

$$\|s_k\| \leq \max\left\{\left(\frac{3(2+\alpha)L_g}{4\sigma_{\min}}\right)^{\frac{1}{\alpha}}, \left(\frac{3(2+\alpha)\|g_k\|}{\sigma_{\min}}\right)^{\frac{1}{1+\alpha}}\right\}, \quad k \geq 0. \tag{2.26}$$

The assumptions A.α, that we employ the true Hessians rather than approximations and that the model is minimized globally imply that the $\alpha \leq 1$ analog of [1, Corollary 2.6] holds, which gives $\|g_k\| \to 0$ as $k \to \infty$, and so $\{\|g_k\|\}$, $k \geq 0$, is bounded above. The bound (2.4) now follows from (2.26).

Using the same techniques as in [1, Lemma 5.2] that applies when f satisfies A.1, it is easy to show for the more general A.α case that $\sigma_k \leq \sigma_{\max}$ for all k, where $\sigma_{\max}$ is defined just after (2.25). It follows from (2.23) that (2.9) holds. Lemma 2.1 now provides that (2.5) is satisfied. The bound (2.25) follows similarly to [2, Corollary 5.3]. □

We cannot extend this result to the $\alpha = 0$ case unless we also assume that H_k is positive semi-definite. If this is the case, we may remove the first term in the max in (2.26), and the remainder of the proof is valid.

We note that bounding the regularization parameter σ_k away from zero in (2.24) appears crucial when establishing the bounds (2.4) and (2.5). Requiring (2.24) implies that the Newton step is always perturbed, but does not prevent local quadratic convergence of ARC [2] and yields improved global worst-case complexity for ARC as can be seen by letting $\alpha = 1$ in (2.25). This ARC bound is better than the global bounds for the steepest-descent and Newton's methods [3]. In Section 3, we show that the bound (2.25) is essentially tight, and that any method in M.α when applied to functions in A.α takes at least $\lfloor \epsilon^{-(2+\alpha)/(1+\alpha)} \rfloor$ function evaluations. Thus from a worst-case complexity point of view, $(2+\alpha)-$regularization methods are the optimal M.α-methods for functions in A.α.

Goldfeld-Quandt-Trotter-type (GQT) methods [11]. Let $\alpha \in (0, 1]$. These algorithms set

$$\lambda_k = \begin{cases} 0, & \text{when } \lambda_{\min}(H_k) \geq \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}; \\ -\lambda_{\min}(H_k) + \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}, & \text{otherwise,} \end{cases} \tag{2.27}$$

in (2.2), where $\omega_k > 0$ is an algorithm parameter that is adjusted so as to ensure sufficient objective decrease. (Observe that replacing $\frac{\alpha}{1+\alpha}$ by 1 in the exponent of $\|g_k\|$ in (2.27) recovers the original method of Goldfeld et al. [11].) It is straightforward to check that (2.3) holds for

the choice (2.27). Thus the GQT approach takes the pure Newton step whenever the Hessian is locally sufficiently positive definite, and a suitable regularization of this step otherwise. The parameter ω_k is increased by a factor, say $\gamma_1 > 1$, and x_{k+1} left as x_k whenever the step s_k does not give sufficient decrease in f (i.e., iteration k is unsuccessful), namely when

$$\rho_k \stackrel{\rm def}{=} \frac{f_k - f(x_k + s_k)}{f_k - m_k(s_k)} \leq \eta_1, \tag{2.28}$$

where $\eta_1 \in (0, 1)$ and

$$m_k(s) = f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \tag{2.29}$$

is the model (2.6) with $\beta_k = 0$. If $\rho_k > \eta_1$, then $\omega_{k+1} \leq \omega_k$ and x_{k+1} is constructed as in (2.1). Similarly to regularization methods, we can bound the total number of unsuccessful iterations as a constant multiple of the successful ones, provided ω_k is chosen such that

$$\omega_k \geq \omega_{\min}, \quad k \geq 0. \tag{2.30}$$

Note that the choice (2.27) implies that (2.5) holds, provided ω_k is uniformly bounded above. We show that the latter, as well as (2.4), hold for functions in A.α.

Lemma 2.5. Let f satisfy A.α with $\alpha \in (0, 1]$. Consider minimizing f by applying a GQT method that sets λ_k in (2.2) according to (2.27), measures progress according to (2.28), and chooses the parameter ω_k to satisfy (2.30). Then (2.4) and (2.5) hold, and so the GQT method belongs to M.α.

Proof. Let us first show (2.4). Since $\omega_k > 0$, and $g_k \neq 0$ until termination, the choice of λ_k in (2.27) implies that $\lambda_k + \lambda_{\min}(H_k) > 0$, for all k, and so (2.2) provides

$$s_k = -(H_k + \lambda_k I)^{-1} g_k,$$

and hence,

$$\|s_k\| \leq \|(H_k + \lambda_k I)^{-1}\| \cdot \|g_k\| = \frac{\|g_k\|}{\lambda_k + \lambda_{\min}(H_k)}, \quad k \geq 0. \tag{2.31}$$

It follows from (2.27) and (2.30) that

$$\lambda_k + \lambda_{\min}(H_k) \geq \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}} \geq \omega_{\min} \|g_k\|^{\frac{\alpha}{1+\alpha}}, \text{ for all } k \geq 0.$$

This and (2.31) further give

$$\|s_k\| \leq \frac{\|g_k\|^{\frac{1}{1+\alpha}}}{\omega_{\min}}, \quad k \geq 0. \tag{2.32}$$

As global convergence assumptions are satisfied when f in A.α [9, 11], we have $\|g_k\| \to 0$ as $k \to \infty$ (in fact, we only need the gradients $\{g_k\}$ to be bounded). Thus (2.32) implies (2.4).

Due to (2.27), (2.5) holds if we show that $\{\omega_k\}$ is uniformly bounded above. For this, we first need to estimate the model decrease. Taking the inner product of (2.2) with s_k, we deduce

$$-g_k^T s_k = s_k^T H_k s_k + \lambda_k \|s_k\|^2.$$

Substituting this into the model decrease, we deduce also from (2.6) with $\beta_k = 0$ that

$$f_k - m_k(s_k) = -g_k^T s_k - \tfrac{1}{2} s_k^T H_k s_k = \tfrac{1}{2} s_k^T H_k s_k + \lambda_k \|s_k\|^2 \geq (\tfrac{1}{2}\lambda_{\min}(H_k) + \lambda_k)\, \|s_k\|^2.$$

It is straightforward to check that this and (2.27) now imply

$$f_k - m_k(s_k) \geq \tfrac{1}{2}\omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}} \cdot \|s_k\|^2. \tag{2.33}$$

We show next that iteration k is successful for ω_k sufficiently large. From (2.28) and second-order Taylor expansion of $f(x_k + s_k)$, we deduce

$$|\rho_k - 1| = \left| \frac{f(x_k + s_k) - m_k(s_k)}{f_k - m_k(s_k)} \right| \leq \frac{|H_k - H(\xi_k)| \cdot \|s_k\|^2}{2(f_k - m_k(s_k))} \leq \frac{L_{H,\alpha} \|s_k\|^{2+\alpha}}{2(f_k - m_k(s_k))}.$$

This and (2.33) now give

$$|\rho_k - 1| \leq \frac{L_{H,\alpha} \|s_k\|^{\alpha}}{\omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}} \leq \frac{L_{H,\alpha}}{\omega_{\min}^{\alpha} \omega_k}, \tag{2.34}$$

where to obtain the last inequality, we used (2.32). Due to (2.28), iteration k is successful when $|\rho_k - 1| \leq 1 - \eta_1$, which from (2.34) is guaranteed to hold whenever $\omega_k \geq \frac{L_{H,\alpha}}{\omega_{\min}^{\alpha}(1-\eta_1)}$. As on each successful iteration we set $\omega_{k+1} \leq \omega_k$, it follows that

$$\omega_k \leq \overline{\omega} \stackrel{\text{def}}{=} \max\left\{ \omega_0, \frac{\gamma_1 L_{H,\alpha}}{\omega_{\min}^{\alpha}(1 - \eta_1)} \right\}, \quad k \geq 0, \tag{2.35}$$

where the max term addresses the situation at the starting point and the γ_1 factor is included in case an iteration was unsuccessful and close to the bound. This concludes proving (2.5). □

Regarding upper bounds for the function-evaluation complexity of GQT methods when applied to functions in A.α, one can show (using the model decrease (2.33) and a lower bound on the step) that GQT takes at most $\mathcal{O}\left(\epsilon^{-\frac{\alpha}{1+\alpha}-2}\right)$, which is worse than the steepest-descent count. Note that this bound improves if only Newton steps are taken, to be of the order $\epsilon^{-\frac{2+\alpha}{1+\alpha}}$; however, this is uncommon for general nonconvex functions. Here, we show in Section 3 that GQT methods are not optimal from a worst-case complexity point of view when applied to A.α objectives.

Trust-region algorithms [9]. These methods compute the correction s_k as the global solution of the subproblem

$$\text{minimize } \; f_k + g_k^T s + \tfrac{1}{2} s^T H_k s \; \text{ subject to } \; \|s\| \leq \Delta_k, \tag{2.36}$$

where Δ_k is an evolving trust-region radius that is chosen to ensure sufficient decrease of f at $x_k + s_k$. The resulting global minimizer satisfies (2.2)–(2.3) [9, Corollary 7.2.2]. The scalar λ_k in (2.2) is the Lagrange multiplier of the trust-region constraint, satisfies

$$\lambda_k \geq \max\{0, -\lambda_{\min}(H_k)\} \tag{2.37}$$

and is such that $\lambda_k = 0$ whenever $\|s_k\| < \Delta_k$ (and then, s_k is the Newton step) or calculated using (2.2) to ensure that $\|s_k\| = \Delta_k$. The scalar $\beta_k = 0$ in (2.6). The iterates are defined

by (2.1) whenever sufficient progress can be made in some relative function decrease (so-called *successful iterations*), and they remain unchanged otherwise (*unsuccessful iterations*) while Δ_k is adjusted to improve the model (decreased on unsuccessful iterations, possibly increased on successful ones). The total number of unsuccessful iterations is bounded above by a constant multiple of the successful ones [12, page 23] provided Δ_k is not increased too fast on successful iterations. One successful iteration requires one gradient and one function evaluation while an unsuccessful one only evaluates the objective.

The property (2.4) of M.α methods can be easily shown for trust-region methods, see Lemma 2.6 below. It is unclear however, whether conditions (2.5) or (2.9) can be guaranteed in general for functions in A.α. The next lemma gives conditions for which a uniform upper bound on the multiplier λ_k can be guaranteed.

Lemma 2.6. Let f satisfy assumptions A.0. Consider minimizing f by applying a trust-region method as described in [9, Algorithm 6.1.1], where the trust-region subproblem is minimized globally to compute s_k and where the trust-region radius is chosen to ensure that

$$\Delta_k \leq \Delta_{\max}, \quad k \geq 0, \tag{2.38}$$

for some $\Delta_{\max} > 0$. Then (2.4) holds. Additionally, if

$$\|g_{k+1}\| \leq \|g_k\|, \text{ for all } k \text{ sufficiently large,} \tag{2.39}$$

then $\lambda_k \leq \lambda_{\max}$ for all k and some $\lambda_{\max} > 0$.

Proof. Consider he basic trust-region algorithm as described in [9, Algorithm 6.1.1], using the same notation. Since the global minimizer s_k of the trust-region subproblem is feasible with respect to the trust-region constraint, we have $\|s_k\| \leq \Delta_k$, and so (2.4) follows trivially from (2.38).

Clearly, the upper bound on λ_k holds whenever $\lambda_k = 0$ or $\lambda_k = -\lambda_{\min}(H_k) \leq L_g$. Thus it is sufficient to consider the case when $\lambda_k > 0$ and $H_k + \lambda_k I \succ 0$. The first condition implies that the trust-region constraint is active, namely $\|s_k\| = \Delta_k$ [9, Corollary 7.2.2]. The second condition together with (2.2) implies, as in the proof of Lemma 2.1, that (2.10) holds. Thus we deduce

$$\Delta_k \leq \frac{\|g_k\|}{\lambda_k + \lambda_{\min}(H_k)},$$

or equivalently,

$$\lambda_k \leq \frac{\|g_k\|}{\Delta_k} - \lambda_{\min}(H_k) \leq \frac{\|g_k\|}{\Delta_k} + L_g, \quad k \geq 0.$$

It remains to show that

$$\{\|g_k\|/\Delta_k\} \text{ is bounded above independently of } k. \tag{2.40}$$

By [9, Theorem 6.4.2], we have that there exists $c \in (0, 1)$ such that the implication holds

$$\Delta_k \leq c\|g_k\| \quad \Longrightarrow \quad \Delta_{k+1} \geq \Delta_k, \text{ i.e., } k \text{ is successful.} \tag{2.41}$$

(Observe that the Cauchy model decrease condition [9, Theorem 6.3.3] is sufficient to obtain the above implication.) Letting $\gamma_1 \in (0,1)$ denote the largest factor we allow Δ_k to be decreased by (during unsuccessful iterations), we will show that

$$\Delta_k \geq \min\{\Delta_{k_0}, c\gamma_1\|g_k\|\} \quad \text{for all } k \geq k_0, \tag{2.42}$$

where k_0 is the iteration from which onwards (2.39) holds; note that since g_k remains unchanged on unsuccessful iterations, (2.39) trivially holds on such iterations. Since the assumptions of [9, Theorem 6.4.6] are satisfied, we have that $\|g_k\| \to 0$, as $k \to \infty$. This and (2.42) imply (2.40). Thus it remains to show (2.42). Using a similar argument to that of [9, Theorem 6.4.3], we let $k \geq k_0$ be the first iterate such that $\Delta_{k+1} < c\gamma_1\|g_{k+1}\|$. Then since $\Delta_{k+1} \geq \gamma_1\Delta_k$ and from (2.39) we have that $\Delta_k < c\|g_k\|$. This and (2.41) give

$$\Delta_{k+1} \geq \Delta_k \geq c\gamma_1\|g_k\| \geq c\gamma_1\|g_{k+1}\|,$$

where to obtain the second and third inequalities, we used the hypothesis and (2.39), respectively. We have reached a contradiction with our assumption that $k+1$ is the first iteration greater than k_0 such that the lower bound on Δ_k does not hold. □

Note that if (2.17) holds for some $\alpha \in [0,1]$, then (2.39) is satisfied, and so Lemma 2.6 shows that if (2.17) holds, then (2.16) is satisfied. It follows from Lemma 2.3 that fast convergence of trust-region methods for functions in A.α alone is sufficient to ensure (2.13), which in turn is connected to our definition of the class M.α. However, the properties of the multipliers (in the sense of (2.5) for any $\alpha \in [0,1]$ or even (2.13)) remain unclear in the absence of fast convergence of the method. Some impractical rules can be constructed that ensure λ_k satisfies (2.13), but at the expense of the resulting trust-region method essentially resembling cubic or other regularization methods. Based on our experience, we are inclined to believe that generally, the multipliers λ_k are at best guaranteed to be uniformly bounded above, even for specialized, potentially computationally expensive, rules of choosing the trust-region radius.

Trust-region methods (with simply a steepest-descent-like Cauchy condition imposed on the step) can be shown to take at most $\mathcal{O}(\epsilon^{-2})$ function evaluations when applied to functions in A.0 [12]. As the Newton step is taken in the trust-region framework satisfying (2.2) whenever it is within the trust region and gives sufficient decrease in the presence of local convexity, the A.1- (hence A.0-) example of inefficient behaviour for Newton's method of complexity precisely ϵ^{-2} can be shown to apply also to trust-region methods [3]. Here, we show that the function-evaluation complexity of trust-region can vary in "inefficiency" between ϵ^{-2} and $\epsilon^{-3/2}$ depending on the smoothness of the objective and on the assumptions on the multipliers.

Linesearch methods with Armijo-Goldstein linesearch [10, 18]. We have not considered here using a linesearch to control improvement in the objective at each step. Such methods compute $x_{k+1} = x_k + \theta_k s_k$, $k \geq 0$, where s_k is defined via a variant

$$(H_k + M_k)s_k = -g_k$$

of (2.2) in which M_k is chosen so that $H_k + M_k$ is "sufficiently" positive definite, and the stepsize θ_k is calculated so as to decrease f (the *linesearch*); this is always possible for sufficiently small θ_k. A change to the definitions (2.1) and (2.2) to include a stepsize that ensures sufficient decrease on each iteration is possible without the examples and analysis that follow changing substantially, but is rather unwieldy. A suitable linesearch to this end is the Goldstein-Armijo technique [18, 3].

3 Examples of inefficient behaviour for methods in M.α

Let $\alpha \in [0,1]$. Our intent is to show that for every method in M.α, we can construct sequences $\{f_k\}$, $\{g_k\}$, $\{H_k\}$ and a function $f^{M.\alpha}(x)$ satisfying A.α such that

$$\|g_k\| \geq \left(\frac{1}{k+1}\right)^{\frac{1+\alpha}{2+\alpha-\tau(1+\alpha)}}, \quad k \geq 0, \text{ for some arbitrarily small } \tau > 0 \ ; \tag{3.1}$$

$$f^{M.\alpha}(x_k) = f_k, \quad \nabla_x f^{M.\alpha}(x_k) = g_k \quad \text{and} \quad \nabla_{xx} f^{M.\alpha}(x_k) = H_k. \tag{3.2}$$

The inequality (3.1) implies that the method takes at least $\lfloor \epsilon^{-\frac{2+\alpha}{1+\alpha}+\tau} \rfloor$ iterations to generate $\|g_k\| \leq \epsilon$, for any $\epsilon > 0$ and for arbitrarily small τ, when applied to minimizing $f^{M.\alpha}(x)$ starting at x_0. This shows that the evaluation complexity of $(2+\alpha)$-regularization methods is essentially optimal in the order of ϵ, as their upper bound is of the same order in ϵ as the lower bound given by our examples; see (2.25).

We consider a one-dimensional example. Assume for now the more general expression for the sequence $\{g_k\}$, namely,

$$g_k = -\left(\frac{1}{k+1}\right)^t, \quad k \geq 0, \text{ for some } t \in (0,1], \tag{3.3}$$

and also, in keeping with the definition of the methods in M.α, that

$$x_0 = 0, \quad x_{k+1} - x_k = s_k = -\frac{g_k}{H_k + \lambda_k}, \text{ with } \lambda_k \geq 0, \tag{3.4}$$

and

$$0 < H_k + \lambda_k \leq \overline{\kappa}_\lambda |g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0, \tag{3.5}$$

for some $\overline{\kappa}_\lambda > 0$ independent of k — a complete justification as to why (3.5) is achieved by methods in M.α when applied to our constructed $f^{M.\alpha}$ is given in the proof of Theorem 3.3.

It follows from (3.3) and (3.4) that

$$s_k > 0 \text{ and } x_k = \sum_{i=0}^{k-1} s_i, \quad k \geq 0. \tag{3.6}$$

We use Hermite interpolation to obtain $f^{M.\alpha}$, namely

$$f^{M.\alpha}(x) = p_k(x - x_k) + f_{k+1} \text{ for } x \in [x_k, x_{k+1}] \text{ and } k \geq 0, \tag{3.7}$$

where p_k is the polynomial

$$p_k(s) = c_{0,k} + c_{1,k}s + c_{2,k}s^2 + c_{3,k}s^3 + c_{4,k}s^4 + c_{5,k}s^5,$$

with coefficients defined by the interpolation conditions

$$\begin{aligned} &p_k(0) = f_k - f_{k+1}, \quad p_k(s_k) = 0; \\ &p_k'(0) = g_k, \quad p_k'(s_k) = g_{k+1}; \\ &p_k''(0) = H_k, \quad p_k''(s_k) = H_{k+1}, \end{aligned} \tag{3.8}$$

where s_k is defined in (3.4). These conditions yield the following values for the coefficients

$$c_{0,k} = f_k - f_{k+1}, \quad c_{1,k} = g_k, \quad c_{2,k} = \tfrac{1}{2}H_k;$$

with the remaining coefficients satisfying

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} \Delta f_k - g_k s_k - \frac{1}{2} s_k^T H_k s_k \\ \Delta g_k - H_k s_k \\ \Delta H_k \end{pmatrix},$$

where

$$\Delta f_k = f_{k+1} - f_k, \quad \Delta g_k = g_{k+1} - g_k \text{ and } \Delta H_k = H_{k+1} - H_k.$$

Hence we obtain, also from (3.4),

$$c_{3,k} = 10\frac{\Delta f_k}{s_k^3} - 4\frac{\Delta g_k}{s_k^2} + \frac{\Delta H_k}{2s_k} - 10\frac{g_k}{s_k^2} - \frac{H_k}{s_k} = 10\frac{\Delta f_k}{s_k^3} - 4\frac{\Delta g_k}{s_k^2} + \frac{\Delta H_k}{2s_k} - 9\frac{g_k}{s_k^2} + \frac{\lambda_k}{s_k};$$
$$c_{4,k} = -15\frac{\Delta f_k}{s_k^4} + 7\frac{\Delta g_k}{s_k^3} - \frac{\Delta H_k}{s_k^2} + 15\frac{g_k}{s_k^3} + \frac{H_k}{2s_k^2} = -15\frac{\Delta f_k}{s_k^4} + 7\frac{\Delta g_k}{s_k^3} - \frac{\Delta H_k}{s_k^2} + \frac{29}{2}\cdot\frac{g_k}{s_k^3} - \frac{\lambda_k}{2s_k^2};$$
$$c_{5,k} = 6\frac{\Delta f_k}{s_k^5} - 3\frac{\Delta g_k}{s_k^4} + \frac{\Delta H_k}{2s_k^3} - 6\frac{g_k}{s_k^4}.$$

To show that $f^{M.\alpha}$ satisfies A.α, recall that $s_k > 0$ due to (3.6), and so (3.7) provides that $f^{M.\alpha}$ is twice continuously differentiable on the nonnegative reals (and it can be extended by continuity to the negative reals). It remains to investigate the gradient's Lipschitz continuity and Hessian's $\alpha-$Hölder continuity, as well as whether $f^{M.\alpha}$ is bounded below. We ensure the remaining properties by further specifying the choice of f_k, g_k and H_k.

Lemma 3.1. Consider an objective $f^{M.\alpha}$ that satisfies (3.3)–(3.5). Let the Hessian of $f^{M.\alpha}$ at x_k be chosen to satisfy

$$\overline{\kappa}_h |g_k|^{\frac{\alpha}{1+\alpha}} \geq H_k \geq -\kappa_h |g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0, \tag{3.9}$$

for some positive constants $\overline{\kappa}_h$ and κ_h independent of k. Then

$$s_k \geq \overline{\kappa}_{s,\alpha} |g_k|^{\frac{1}{1+\alpha}}, \quad k \geq 0, \tag{3.10}$$

for some $\overline{\kappa}_{s,\alpha} > 0$, and

$$\left|\frac{g_k}{s_k^{1+\alpha}}\right|, \quad \left|\frac{H_k}{s_k^{\alpha}}\right| \text{ and } \left|\frac{\lambda_k}{s_k^{\alpha}}\right| \text{ are bounded above, independently of } k. \tag{3.11}$$

Additionally, if

$$s_k \leq \overline{\kappa}_s, \quad k \geq 0, \tag{3.12}$$

and

$$|f_k - f_{k+1}| \leq \kappa_f s_k^{2+\alpha}, \quad k \geq 0, \tag{3.13}$$

for some $\overline{\kappa}_s > 0$ and $\kappa_f > 0$ independent of k, then the gradient of $f^{M.\alpha}$ is globally Lipschitz continuous and the Hessian of $f^{M.\alpha}$ is globally $\alpha-$Hölder continuous along the path of the iterates. Finally, if

$$|f_k| \leq \overline{\kappa}_f, \quad k \geq 0, \ \text{ for some } \overline{\kappa}_f > 0, \tag{3.14}$$

then $f^{M.\alpha}$ is bounded below (on the nonnegative reals).

Proof. From (3.4), we have $s_k = |g_k|/(H_k + \lambda_k)$ and so (3.5) provides (3.10) with $\overline{\kappa}_{s,\alpha} \stackrel{\rm def}{=} 1/\overline{\kappa}_\lambda$. Thus the first ratio in (3.11) is uniformly bounded above. The uniform boundedness of the second ratio in (3.11) follows from (3.9) and (3.10). Now we deduce from (3.5) and (3.9) that

$$0 \leq \lambda_k \leq (\overline{\kappa}_\lambda + \kappa_h)|g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0, \tag{3.15}$$

which together with (3.10), provides that the third ratio in (3.11) is uniformly bounded above.

We next show that the Hessian of $f^{M.\alpha}$ is globally $\alpha-$Hölder continuous on the path of the iterates, namely that (2.8) holds. From (3.7), this is implied by

$$|p^{'''}(s)| \leq c|s|^{-1+\alpha}, \ \text{ for all } s \in [0, s_k] \text{ and for some } c > 0 \text{ independent of } s \text{ and } k. \tag{3.16}$$

We have from the expression of p_k and $s \in [0, s_k]$ that

$$|p^{'''}(s)| \cdot |s|^{1-\alpha} \leq (6|c_{3,k}| + 24|c_{4,k}|s_k + 60|c_{5,k}|s_k^2)s_k^{1-\alpha} = 6|c_{3,k}|s_k^{1-\alpha} + 24|c_{4,k}|s_k^{2-\alpha} + 60|c_{5,k}|s_k^{3-\alpha}. \tag{3.17}$$

It follows from the expressions of the coefficients of p_k that the right-hand side of (3.17) is bounded above provided the terms

$$\left|\frac{\Delta f_k}{s_k^{2+\alpha}}\right|, \left|\frac{\Delta g_k}{s_k^{1+\alpha}}\right|, \left|\frac{\Delta H_k}{s_k^{\alpha}}\right|, \left|\frac{g_k}{s_k^{1+\alpha}}\right|, \left|\frac{\lambda_k}{s_k^{\alpha}}\right|, \tag{3.18}$$

are uniformly bounded above independently of k. Clearly, the first expression follows from (3.13), while the remaining ones, from (3.11) and the expression of g_k in (3.3).

To show that the gradient of f^M is globally Lipschitz continuous is equivalent to proving that $p_k^{''}(s)$ is uniformly bounded above on the interval $[0, s_k]$. Since $s_k > 0$, we have

$$|p_k^{''}(s)| \leq 2|c_{2,k}| + 6|c_{3,k}|s_k + 12|c_{4,k}|s_k^2 + 20|c_{5,k}|s_k^3, \quad s \in [0, s_k].$$

The above explicit expressions of the coefficients of p_k imply that it is enough to show that the quantities

$$\left|\frac{\Delta f_k}{s_k^2}\right|, \left|\frac{\Delta g_k}{s_k}\right|, |\Delta H_k|, \left|\frac{g_k}{s_k}\right| \text{ and } |\lambda_k| \tag{3.19}$$

are uniformly bounded above, independently of k. But all ratios in (3.19) can be expressed as the corresponding ratios in (3.18) multiplied by s_k^{α}, while s_k is bounded above due to (3.12). Hence (3.18) implies that the ratios in (3.19) are uniformly bounded above.

It remains to show that $f^{M.\alpha}$ is bounded below, which due to (3.7) and (3.14), is equivalent to proving that $|p_k(s)|$ is uniformly bounded above for $s \in [0, s_k]$. Recalling the expressions of the coefficients of p_k, and s_k being bounded above from (3.12), this now follows from $\{f_k - f_{k+1}\}$, $\{|g_k|\}$, $\{H_k\}$ and $\{\lambda_k\}$ being bounded above due to (3.13) and (3.11). □

Note that we have shown that $f^{M.\alpha}(x)$ is bounded below for $x \geq 0$, which is the domain of interest since $x_k \geq 0$; we can extend $f^{M.\alpha}$ by continuity for $x < 0$ [3]. Note also that though $f^{M.\alpha}$ is bounded below, we have not shown that $f^{M.\alpha}$ is bounded below by the limit, or a lower bound, of the sequence $\{f_k\}$; the latter will often hold as can be seen from the examples in [3].

Clearly, from Lemma 3.1, in order to complete our construction of a suitable function $f^{M.\alpha}$, we need to find suitable choices of $\{f_k\}$ such that (3.13) and (3.14) hold, and in the same vein, to ensure that the function values $\{f_k\}$ are not only monotonically decreasing but that a sufficient decrease in f is gained from x_k to x_{k+1} so that progress towards a minimum is made with each step. The next lemma addresses these issues by making use of the local model (2.6); condition (3.12) will be easily satisfied later on, especially as we are looking to make s_k small in order to capture the worst-case behaviour of the methods.

Lemma 3.2. Consider an objective $f^{M.\alpha}$ that satisfies (3.3)–(3.5) and (3.9). Let also the values f_{k+1} be chosen recursively from f_k so as to satisfy

$$f_{k+1} = f_k - \eta\,(f_k - m_k(s_k))\,, \quad \text{for some } \eta > 0 \text{ independent of } k, \tag{3.20}$$

where $m_k(s)$ is defined in (2.6). Then

$$f_k > f_{k+1}, \quad \text{for all } k. \tag{3.21}$$

Furthermore, if (3.12) holds and

$$f_0 - f_{\text{low}} \leq \sum_{k=0}^{\infty} (f_k - f_{k+1}) < \infty, \tag{3.22}$$

where f_{low} is some lower bound on $\{f_k\}$, then $f^{M.\alpha}$ belongs to A.α.

Proof. It follows from (3.20) and (3.4) that

$$f_k - f_{k+1} = -\frac{\eta}{2} g_k s_k + \frac{\eta}{2} s_k^2 (\lambda_k - \beta_k). \tag{3.23}$$

Now (3.21) follows from $g_k < 0$ due to (3.3), $s_k > 0$, and $\lambda_k \geq \beta_k$ due to (2.6). Relation (3.23), (2.6) and the Cauchy-Schwarz inequality imply that

$$f_k - f_{k+1} \leq \eta |g_k| \cdot s_k + \eta \lambda_k s_k^2, \quad k \geq 0,$$

and furthermore,

$$\frac{f_k - f_{k+1}}{s_k^{2+\alpha}} \leq \eta \frac{|g_k|}{s_k^{1+\alpha}} + \eta \frac{\lambda_k}{s_k^{\alpha}}. \tag{3.24}$$

Since the conditions of Lemma 3.1 are satisfied, (3.11) holds, which together with (3.24), implies that (3.13) is satisfied. Clearly, (3.21) implies (3.14) holds whenever (3.22) is achieved. Lemma 3.1 now provides that $f^{M.\alpha}$ is in A.α. □

By specifying the recursion (3.20) that generates the values f_{k+1} from f_k so that f_{k+1} is in (absolute or relative) agreement with the local model of the function, we have ensured that sufficient progress is made on each iteration towards the solution; indeed, condition (3.20) is the common positive relative decrease requirement for updating the step in trust-region, regularization and other second-order methods. Finally, ensuring (3.12) and (3.22) will be a by-product of the potentially slow rate of convergence of the methods in M.α, which we address next.

Recall that whenever a method in M.α is applied to minimizing a function in A.α, it generates steps that satisfy (2.13); see Lemma 2.2. For the function $f^{M.\alpha}$ constructed in Lemmas 3.1 and 3.2, this reduces to s_k satisfying (3.10). Since the smaller the step the slower the method, it follows that the method will be slowest when (3.10) holds with equality, or equivalently, when s_k varies as $|g_k|^{\frac{1}{1+\alpha}}$, i.e.,

$$s_k = \Theta\left(|g_k|^{\frac{1}{1+\alpha}}\right), \quad k \geq 0, \tag{3.25}$$

where $\Theta(\cdot)$ denotes the existence of upper and lower bounds of the same order as its argument. Relation (3.25) and (3.13) — the latter holding due to Lemma 3.2 — imply that

$$\sum_{k=0}^{\infty}(f_k - f_{k+1}) < \infty \ \text{ when } \ \sum_{k=0}^{\infty}|g_k|^{\frac{2+\alpha}{1+\alpha}} < \infty. \tag{3.26}$$

(Note that the first series in (3.26) is equivalent to (3.22).) Due to (3.3) and the properties of the Riemann zeta function, we have

$$\sum_{k=0}^{\infty}|g_k|^{\frac{2+\alpha}{1+\alpha}} < \infty \quad \Longleftrightarrow \quad \sum_{k=0}^{\infty}\left(\frac{1}{k+1}\right)^{t\cdot\frac{2+\alpha}{1+\alpha}} < \infty \quad \Longleftrightarrow \quad \frac{1+\alpha}{2+\alpha} < t \leq 1. \tag{3.27}$$

Thus, for worst complexity, t in the expression of g_k in (3.3) must be arbitrarily close to the lower bound of the interval in (3.27), and hence of the form

$$t = \frac{1+\alpha}{2+\alpha} + \delta, \ \text{ for some arbitrarily small } \delta > 0. \tag{3.28}$$

We are therefore left with arguing that the choice (3.25) can indeed happen. Since g_k is prescribed, small steps s_k, namely (3.25), are equivalent to $H_k + \lambda_k$ in (3.4) being as large as possible, namely, of the same order as the right-hand side of (3.5). We can achieve this by further taking up the freedom in the choice (3.9) of H_k, as we show next, in the main theorem of this Section. This theorem completes the construction of the lower complexity bound for M.α and results in the optimality of regularization methods for functions in A.α.

Theorem 3.3. Consider an objective $f^{M.\alpha}$ that satisfies (3.3) with t defined in (3.28), (3.4), (3.5) and (3.20). Let

$$H_k = \overline{\kappa}_h |g_k|^{\frac{\alpha}{1+\alpha}}, \quad k \geq 0, \tag{3.29}$$

for some positive constant $\overline{\kappa}_h$ independent of k. Then $f^{M.\alpha}$ belongs to A.α. In (3.28), set

$$\delta = \frac{\tau(1+\alpha)^2}{(2+\alpha)^2 - \tau(1+\alpha)(2+\alpha)}, \ \text{ for some arbitrarily small } \tau > 0. \tag{3.30}$$

Let a method from the class M.α be applied to minimizing $f^{M.\alpha}$. Then the method will take precisely $\lceil \epsilon^{-\frac{2+\alpha}{1+\alpha}+\tau} \rceil$ iterations and function evaluations to generate $|g_k| \leq \epsilon$. In particular, the $(2+\alpha)$-regularization methods are optimal from a worst-case evaluation complexity point of view for the class of A.α-objectives.

Proof. The choice (3.29) and $\lambda_k \geq 0$ implies that

$$H_k + \lambda_k \geq \overline{\kappa}_h |g_k|^{\frac{\alpha}{1+\alpha}},$$

which together with (3.4) give that

$$s_k \leq \frac{1}{\overline{\kappa}_h} |g_k|^{\frac{1}{1+\alpha}}.$$

This and (3.10) imply (3.25). The choice (3.28) of t implies that (3.22) holds as argued just above the statement of the Theorem. Furthermore, (3.25) and (3.3) provide that (3.12) holds. Now all conditions in Lemma 3.2 have been satisfied, and so $f^{M.\alpha}$ is in A.α.

Consider applying a method in M.α to minimizing $f^{M.\alpha}$, with starting point given in (3.4). Then, clearly the remaining conditions in (3.4) are achieved, and so is (3.5). Indeed, since we are in the univariate case and so $H_k = \lambda_{\min}(H_k)$, (2.5) and (3.29) imply that (3.5) holds with $\overline{\kappa}_\lambda \stackrel{\text{def}}{=} \kappa_\lambda \max\{\overline{\kappa}_h, 1\}$. Condition (2.4) is also achieved on our objective due to (3.12), which holds as we have argued in the previous paragraph.

The iteration and evaluation complexity of any method in M.α applied to $f^{M.\alpha}$ follows from (3.28), (3.30) and the argument following (3.1) and (3.2). In particular, if the method of choice is the $(2+\alpha)-$regularization described in (2.22), which belongs to M.α due to Lemma 2.4, then it satisfies a complexity upper bound of the same order in ϵ, with $\tau = 0$; see (2.25). As the upper and lower bound on $(2+\alpha)-$regularization coincides in the order of ϵ, it is optimal from a worst-case complexity point of view, within the class M.α. □

Note that in Theorem 3.3, we could have derived the value of t in (3.28), rather than assume it. Indeed, recalling the argument just before the statement of the Theorem, if we want the construction of $f^{M.\alpha}$ to be well-defined, namely (3.22) to hold, we must have that t satisfies the last relation in (3.27). Furthermore, the smaller the value of t, the worst the complexity of the method, and so we set t to the value in (3.28).

Finite minimizers. Note that the choice of s_k in Theorem 3.3 which satisfies (3.25) implies, due also to (3.3), (3.28) and the properties of the Riemann zeta function, that

$$\sum_{k=0}^{\infty} s_k = \infty.$$

Thus the minimizer/stationary point of our examples is at infinity. However, at termination, the iterate x_k with $|g_k| \leq \epsilon$ is finite for any $\epsilon > 0$, and so $f^{M.\alpha}$ can be extended smoothly beyond x_k in such a way that the resulting function has a unique, finite and global minimizer. Thus, by fixing the required accuracy ϵ and using it in the construction of the objective, we obtain similar examples of inefficiency, with the same complexity for problems with finite minimizers. □

3.1 Illustrations

Let us illustrate the examples of Section 3 for specific methods in M.α. In particular, once we know the choice of method in M.α, there is more freedom in the choice of examples than

prescribed by Theorem 3.3, namely in the choice of H_k, and we can describe the examples in a more method-dependent way.

Let $\alpha \in [0,1]$. Assume g_k is defined as in (3.3) with the choice of t given in (3.28) and (3.30), and hence

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{1+\alpha}{2+\alpha}+\delta}, \tag{3.31}$$

and let H_k satisfy (3.9). Let f_k be defined recursively using (3.20). Let

$$s_k = \left(\frac{1}{k+1}\right)^q, \quad \text{for some } q \in (0,1]. \tag{3.32}$$

It follows from (3.10) that

$$q \leq \frac{1}{2+\alpha} + \frac{\delta}{1+\alpha}. \tag{3.33}$$

Furthermore, for (3.25) to hold, we must have

$$q > \frac{1}{2+\alpha}. \tag{3.34}$$

Hence from (3.33) and (3.34), we have $q \in \left(\frac{1}{2+\alpha}, \frac{1}{2+\alpha} + \frac{\delta}{1+\alpha}\right]$, and since $\delta > 0$ can be arbitrarily small, we must settle for

$$q = \frac{1}{2+\alpha} + \frac{\delta}{1+\alpha}. \tag{3.35}$$

Note that from (3.4) and (3.35), we have

$$H_k + \lambda_k = \left(\frac{1}{k+1}\right)^{\alpha\left(\frac{1}{2+\alpha}+\frac{\delta}{1+\alpha}\right)} \tag{3.36}$$

and (2.5) and (3.5) are verified. Now consider the particular values of λ_k and β_k for some methods in M.α and specialize the examples for these methods.

Newton's method. Recalling (2.21), we have from (3.36) that the choice

$$H_k = \left(\frac{1}{k+1}\right)^{\alpha\left(\frac{1}{2+\alpha}+\frac{\delta}{1+\alpha}\right)}, \quad k \geq 0, \tag{3.37}$$

in the construction of $f^{M.\alpha}$ will generate the required complexity of order $\epsilon^{-(2+\alpha)/(1+\alpha)+\tau}$ iterations. Note that (3.37) is of the same order as (3.29), and that $H_k > 0$ so that the Newton iteration is well-defined.

Since Newton's method belong to M.α for every $\alpha \in [0,1]$, we conclude from our results here and in [3] that Newton's method may take essentially between at least ϵ^{-2} and $\epsilon^{-3/2}$ evaluations to generate $|g_k| \leq \epsilon$; recall that ϵ^{-2} is the sharp order of complexity of steepest-descent method [3]. Note that if the Hessian of the objective is unbounded, and hence, we are outside of the class A.1, the complexity of Newton's method worsens, and in fact, it may be arbitrarily bad [3].

Cubic and other regularizations. Recalling (2.23), we set the parameter $\sigma_k = \sigma > 0$ for all k in the algorithm, which is allowed as every iteration is successful due to (3.20). From (3.36) and the definition of $\lambda_k = \sigma\|s_k\|^\alpha$ in (2.23), the choice

$$H_k = (1-\sigma)\left(\frac{1}{k+1}\right)^{\alpha\left(\frac{1}{2+\alpha}+\frac{\delta}{1+\alpha}\right)}$$

in the construction of $f^{M.\alpha}$ will generate the required complexity of order $\epsilon^{-(2+\alpha)/(1+\alpha)+\tau}$ iterations, which for the $(2+\alpha)-$regularization method is a tight bound in the order of ϵ for objectives in A.α. Letting $\alpha = 1$, note that for $\sigma = 1$, we get $H_k = 0$ which recovers the choice in the cubic regularization example in [3].

Goldfeld-Quandt-Trotter. Recalling (2.27), we can set $\omega_k = \omega$ in the algorithm as every iteration is sucessful due to (3.20) and $H_k = 0$ in $f^{M.\alpha}$, which gives $\lambda_k = \omega|g_k|^{\frac{\alpha}{1+\alpha}}$, which is in agreement to (3.36), due also to (3.31). This choice for H_k in $f^{M.\alpha}$ yields the complexity of $\epsilon^{-(2+\alpha)/(1+\alpha)+\tau}$ Goldfeld-Quandt-Trotter iterations to drive the gradient below ϵ.

Trust-region methods. Recall the choices (2.37) we make in this case. If $\lambda_k = 0$, the trust-region constraint $\|s\| \leq \Delta_k$ is inactive at s_k, in which case, s_k is the Newton step. If we make precisely the choices we made for Newton's method above, choosing Δ_0 such that $\Delta_0 > |s_0|$ implies that the Newton step will be taken in the first and in all subsequent iterations since each iteration is successful and then Δ_k remains unchanged or increases while the choice (3.32) implies s_k decreases. Thus in this case, the trust-region approach, through the Newton step, has the same complexity when applied to $f^{M.\alpha}$ as the Newton step, namely $\epsilon^{-(2+\alpha)/(1+\alpha)+\tau}$, for any $\alpha \in [0, 1]$.

By contrast when $\lambda_k > 0$ for all k, $s_k = \Delta_k$. Using the notation in [9, Algorithm 6.1.1], let η in (3.20) be equal to η_1, which corresponds to successful but not very successful steps s_k. This allows the trust-region radius Δ_k to decrease slightly, namely $\Delta_{k+1} \in [\gamma_2\Delta_k, \Delta_k]$. We let

$$\Delta_{k+1} = \gamma_k\Delta_k, \quad \text{where } \gamma_k = \left(\tfrac{k+1}{k+2}\right)^q$$

with q defined in (3.35). If γ_2 is chosen such that $\gamma_2 \leq (1/2)^q$, then clearly the above updating rule implies that $\Delta_{k+1} \in [\gamma_2\Delta_k, \Delta_k]$. Note that due to (3.32), this updating rule is consistent with the trust-region constraint being active on each iteration. Now we can choose $H_k < 0$ so as to ensure (3.36), noting that despite not knowing the precise value of λ_k for the trust-region method, we know that the global solution of the trust-region subproblem is unique whenever $H_k + \lambda_k I$ is positive definite, which is clearly the case here, due to (3.36).

4 Conclusions

We have provided lower bounds on the evaluation complexity of second-order methods for reaching approximate first-order critical points of nonconvex, smooth unconstrained optimization problems. We have found that regularization algorithms are optimal from a worst-case complexity point of view within the wide class of methods M.α, in that their upper complexity bounds match in order the lower bound we have shown for relevant, sufficiently smooth objectives A.α. Note that every iteration complexity bound discussed here is of the order ϵ^{-p} (for various values of $p > 0$) for driving the objective's gradient below ϵ; thus the methods we have addressed may require an exponential number of iterations $10^{p\cdot k}$ to generate k correct digits in the solution. Also, as our examples are one-dimensional, they fail to capture the problem-dimension dependence of the upper complexity bounds. Indeed, besides the accuracy tolerance ϵ, existing upper bounds depend on the distance to the solution set, that is $f(x_0) - f_{\text{low}}$, and the gradient's and Hessian's Lipschitz or Hölder constants, all of which may dependent on the problem dimension. Some recent developments in this respect can be found in [14].

The methods we have addressed assume that subproblems are solved to global optimality in each iteration in order to compute the step, thus ensuring best possible decrease locally. As such, approximate variants of the algorithms are unlikely to perform better in the worst case than the exact variants discussed here.

When convexity or strong convexity of the objective is assumed, much is known about upper complexity bounds but little about the lower bounds or worst-case optimality of second order methods; the latter has been fully resolved for first-order methods [15]. A sharp bound for cubic regularization methods in the convex case was given in [5], but it is unknown whether this is a lower bound on the wider class of second-order methods.

Here we have solely addressed the complexity of generating first-order critical points, but it is common to require second-order methods for nonconvex problems to achieve second-order criticality. Indeed, upper complexity bounds are known in this case for cubic regularization and trust-region methods [17, 2, 6], which are sharp in some cases [6]. A lower bound on the whole class of second order methods for achieving second-order optimality remains to be established, especially when different accuracy is requested in the first- and second-order criticality conditions.

Regarding the evaluation complexity of constrained optimization problems, we have shown [4, 7, 8] that the presence of constraints does not change the order of the bound, so that the unconstrained upper bound for some first- or second-order methods carries over to the constrained case; note that this does not include the cost of solving the constrained subproblems as the latter does not require additional problem evaluations. Since constrained problems are at least as difficult as unconstrained ones, these bounds are also tight. It remains an open question whether a unified treatment such as the one given here can be provided for the complexity of methods for constrained problems.

References

[1] C. Cartis, N. I. M. Gould and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[2] C. Cartis, N. I. M. Gould and Ph. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, DOI: 10.1007/s10107-009-0337-y, 2010 (online).

[3] C. Cartis, N. I. M. Gould and Ph. L. Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, 20:2833–2852, 2010.

[4] C. Cartis, N. I. M. Gould and Ph. L. Toint. An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity. ERGO Technical Report 09-004, School of Mathematics, University of Edinburgh, 2009.

[5] C. Cartis, N. I. M. Gould and Ph. L. Toint. Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. ERGO Technical Report 10-006, School of Mathematics, University of Edinburgh, 2010.

[6] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. ERGO Technical Report 10-007, School of Mathematics, University of Edinburgh, 2010. To appear in *Journal of Complexity*.

[7] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming . ERGO Technical Report 11-002, School of Mathematics, University of Edinburgh, 2011.

[8] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of finding first-order critical points in constrained nonlinear programming. ERGO Technical Report 11-005, School of Mathematics, University of Edinburgh, 2011.

[9] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. SIAM, Philadelphia, USA, 2000.

[10] J. E. Dennis and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1983. Reprinted as *Classics in Applied Mathematics 16*, SIAM, Philadelphia, USA, 1996.

[11] S. M. Goldfeld, R. E. Quandt and H. F. Trotter. Maximization by quadratic hill-climbing. *Econometrica*, 34:541–551, 1966.

[12] S. Gratton, A. Sartenaer and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

[13] A. Griewank. The modification of Newton's method for unconstrained optimization by bounding cubic terms. Technical Report NA/12 (1981), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom, 1981.

[14] F. Jarre. On Nesterov's smooth Chebyshev-Rosenbrock function. Technical Report, 2011. Available at *Optimization Online*.

[15] Yu. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.

[16] Yu. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[17] Yu. Nesterov and B. T. Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[18] J. Nocedal and S. J. Wright. *Numerical Optimization*. Second edition, Springer-Verlag, New York, USA, 2006.

[19] S. A. Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.

[20] M. Weiser, P. Deuflhard and B. Erdmann. Affine conjugate adaptive Newton methods for nonlinear elastomechanics. *Optimization Methods and Software*, 22(3):413–431, 2007.